开放获取论文推送转发服务系统 iSwitch：论文分发推送*

钱力 [1,2]　师洪波 [1]　张晓林 [1]　梁娜 [1]

[1]（中国科学院文献情报中心 北京 100190）

[2]（中国科学院大学 北京 100049）

摘要：[目标]现将接收与解析成功的开放获取论文分发推送到作者机构和资助机构知识库。**[方法]** 分析 iSwitch 系统技术框架，设计论文分发推送服务功能模块，利用任务调度代理与 FTP 协议实现论文的分发推送。**[结果]** iSwitch 系统可以实现论文分发推送服务，并完成来自 Web of Science 的 34 332 条文章数据的分发推送。**[局限]** 目前仅仅基于一种数据源完成论文分发推送，对基于多个数据推送方的更大体量数据的分发推送服务中可能遇到的问题考虑不够。**[结论]** 实验表明，分发推送服务的工作流程机制是正确的，分发效率满足未来服务需求。

关键词： 开放获取　iSwitch　分发推送　FTP　机构知识库

分类号：G250　TP393

Router Service Engine iSwitch for Open Access Articles: Pushing and Routing

Qian Li[1,2]　Shi Hongbo[1]　Zhang Xiaolin[1]　Liang Na[1]

[1] (National Science Library, Chinese Academy of Sciences, Beijing 100190,China)

[2] (University of Chinese Academy of Sciences, Beijing 100049,China)

Abstract:

[Objective] Routing Open Access articles which are received and parsed successfully to Institute Repository of Author and Funding. **[Methods]** Analyzing the technology framework of iSwitch, then designing the service architecture and function interface of Routing Articles, and finally using Task agent and FTP and Routing Articles.**[Results]** Based on 34 332 articles from Web of Science, finally which are routed successfully by iSwitch.**[Limitations]** Now only carried out routing based on one data source, but the consideration may not enough that the problems may happen in future service. **[Conclusions]** The experiments show that workflow mechanism of pushing and routing is correct and it's efficiency meet future demand for services.

Keywords: Open Access　iSwitch　Pushing and routing　FTP　IR

1 引言

为了支持科研论文的开放获取，促进研究单位的机构知识库[1]的构建，避免由作者自行存缴而带来的繁琐工作、数据缺失或者版本混乱等问题，建立科研成果有效传播的快速通道和促进科研论文的开放共享，中国科学院文献情报中心作为开放获取[2]运动的有力推动者，于 2014 年提出开放获取论文推送转发服务系统 iSwitch[3]。在此背景下，笔者在《开放获取论文推送转发服务系统 iSwitch：

*本文系中国科学院文献情报能力建设专项“国际开放论文国家交换服务中心示范系统”(项目编号:Y14008)的研究成果之一。

技术流程与标准》[4]中为 iSwitch 系统的建设提出了具体的技术要求与遵循标准，并且根据推送方、转发方与接收方的需求，进行了分发推送服务工作流分析。

在《开放获取论文推送转发服务系统 iSwitch：论文接收与解析》一文中提出构建 iSwitch 系统的接收与解析功能的具体设计方案，分别使用 FTP[5]与 SWORD[6]协议接收出版商推送的开放科技论文，并进行系统实现。根据 iSwitch 系统的基本框架与标准流程，需要构建 iSwitch 系统的推送转发模块，以实现将通过 iSwitch 系统接收、解析成功的科技论文分发推送到相应机构与资助基金的机构知识库。本文即在此需求背景下，按照统一技术规范要求，设计并实现 iSwitch 系统的分发推送功能。

2 iSwitch 论文分发推送技术实现框架

iSwitch 论文分发推送模块基于 FTP 通信协议实现分发推送数据信息的通信传输，具体功能包括：创建分发推送任务；执行分发推送任务，将论文分发到所对应的机构或者基金组织的 FTP 目录下，目标接收机构按照预定的接收周期进行收割；分发推送审计；分发推送信息配置。整体技术实现框架如图 1 所示：

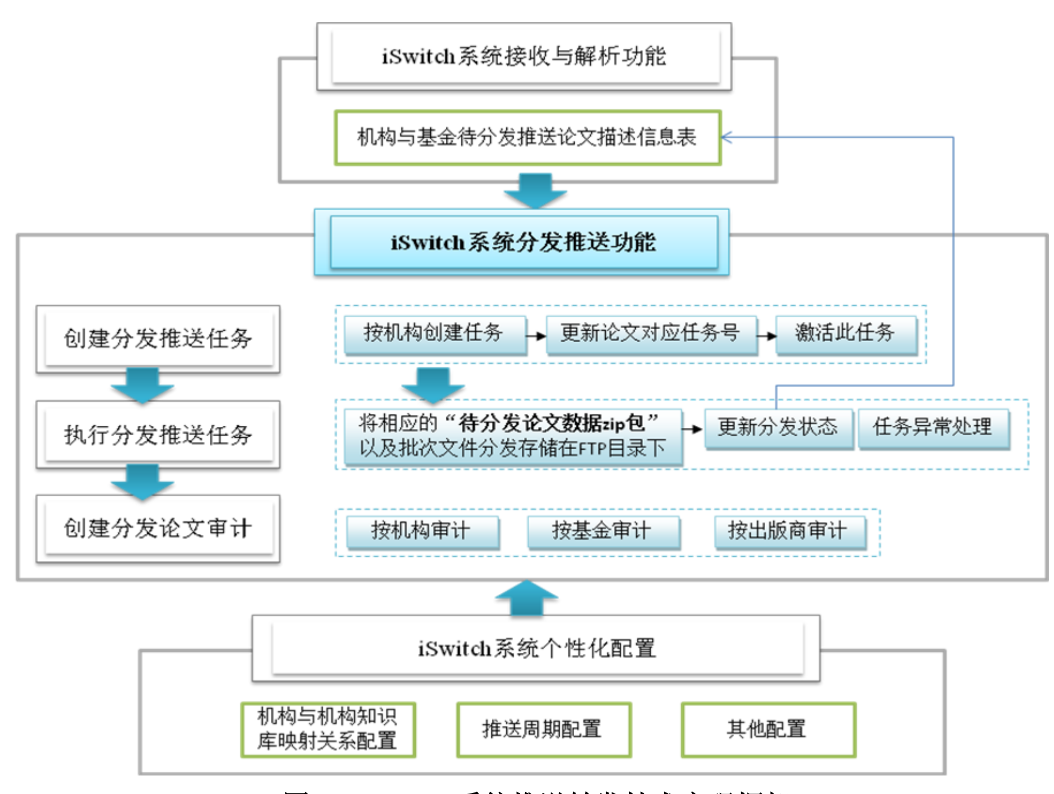


图 1 iSwitch 系统推送转发技术实现框架

2.1 创建分发推送任务

iSwitch 系统设计了自动创建“论文分发推送任务”的生成器(taskCreator)，基于前一天（截止到创建任务时间当天的零点）已经正确解析与映射到相应作者机构与资助基金的开放论文数据列表，自动生成可分发推送任务序列号，同时保存到待分发任务列表中，供执行分发触发器定时调用。

2.2 执行分发推送任务

执行分发推送任务通过 iSwitch“论文分发推送任务”的分发器（taskRouter），将任务列表中处于可分发状态的任务即时分发推送到 FTP 服务器上所对应机构的目录下，等待机构与资助机构按照接收频次定时收割，当收割接收成功后，iSwitch 系统会自动更新所分发开放论文的分发状态（已分发），以备后续的审计，同时也确保同一篇开放论文不在同一个目标机构与资助机构中重复分发推送。其中接收分发论文的频率在 iSwitch 系统中分为天、周、月、季度以及年多个时间窗。

执行分发任务过程中，“待分发论文数据 ZIP 包”是在论文接收与解析部分按照 ZIP 包的格式封装的，内容包括论文全文文件（PDF/XML）、元数据描述 XML 文件以及推送方提供的其他数据文件。其中，元数据描述 XML 文件是按照 JATS[7]标准描述格式完成推送数据论文包的封装。

2.3 分发推送论文审计

分发推送论文审计从数据来源、审计类型（分发任务、作者机构和资助基金）以及审计范围（时间点、时间段）三个维度，审计已经分发推送论文的结果概况，即成功与失败分发的论文记录：

（1）**成功分发**，可以查看分发目标机构、在机构知识库中的 URL 以及资助基金名称；

（2）**失败分发**，系统会自动标识失败分发可能的原因，修改之后 iSwitch 系统自动调用异常处理机制，继续分发推送。

2.4 分发推送信息配置

为了支持推送方（出版商）更方便、规范地向 iSwitch 服务推送论文，同时让接收方（作者机构与资助机构）更可靠、便捷地接收自己单位的科研论文，iSwitch 系统具有灵活的个性化配置策略，其中与分发推送服务相关的主要有以下配置：

（1）**机构与机构知识库映射关系配置**：多个机构合并到一个机构和多个机构共用映射到同一个知识库的配置；

（2）**分发推送周期配置**：根据不同机构知识库接收论文的数量，其接收周期也可以进行个性化配置；

（3）**其他配置**：对于分发机构还需进行知识库 IP、机构知识库网站等机构详细信息配置，这些配置可以起到保证分发的安全性、提供核对信息等作用。

3 iSwitch 论文分发推送模块设计

为了保障 iSwitch 服务系统的论文分发推送模块设计的科学性与合理性，保障分发推送的有效性与高效性，iSwitch 论文分发推送模块设计主要遵循以下 4 个原则：

（1）功能模块之间松耦合与功能单元内部的高内聚特征，方便功能的调用与后期维护；

（2）遵循工作流执行规范，设计与封装相应的功能单元，以服务接口形式对外暴露发布，供 iSwitch 系统调用；

（3）iSwitch 服务系统分发推送功能的自动化执行；

（4）具有异常情况自动处理机制。

根据分发推送整体技术框架，分发推送模块共分为任务定时创建模块、任务即时执行模块、论文分发状态更新模块与异常分发任务自动处理模块共 4 个部分，整体结构如图 2 所示：

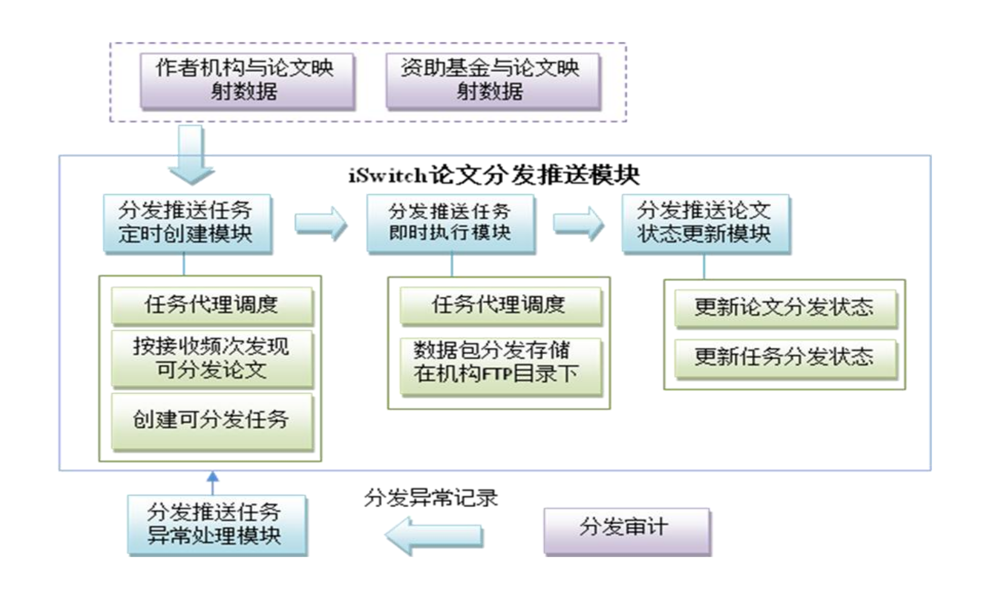


图 2　iSwitch 论文分发推送模块设计结构图

3.1 分发推送论文任务定时创建模块

此模块功能是将已经正确解析、映射到作者机构与资助机构的开放论文，通过代理自动调度的形式，创建与论文相对应可分发的任务，其中需要解决以下关键问题：

（1）任务代理自动调度。为了增强 iSwitch 系统的自动分发功能，分发模块针对任务创建设计与实现了“任务代理调度(Agent)”功能，即每天 iSwitch 会自动调度“任务创建功能”，按照设计的流程自动执行，完成任务的创建。

（2）按机构接收频次获取可分发论文。代理自动调度启动之后，iSwitch 会自动发现今天之前（当前零点之前）的正确解析而且未分发的开放论文，从中获取需要分发的机构或者资助机构列表，将这些机构的论文接收频次与当前时间进行自动计算分析，如果满足接收频次，则此机构未分发的论文即可加入分发任务列表中，否则过滤掉不满足接收频次的未分发论文，等待下次任务的调度。

（3）**创建可分发任务**。分发任务的创建以“按机构接收频次获取可分发论文”功能单元所计算分析的机构 ID 为标识，即一个可分发机构对应一个新的“可分发状态”的任务号 TaskID，同时将此机构所对应的未分发论文映射的任务号更新为 TaskID。

3.2 分发推送任务即时执行模块

此模块功能是将处于可分发状态的任务即时执行（iSwitch 服务配置了每 5 秒钟自动扫描是否有待分发的任务），分发推送论文并存储在作者机构与资助基金所对应的 FTP 目录，具体执行流程是：

（1）找到作者机构；

（2）获取分发批次号码；

（3）获取与分发此批次号下的所有待分发论文数据 ZIP 包。

按照上述的流程路径，将压缩数据包基于 FTP 协议分发存储在服务器空间中，供作者机构定时收割与解析。

3.3 分发推送论文状态更新模块

此模块以 Web Serives 接口形式对外发布，供机构知识库平台调用，以实现更新已经分发推送的论文分发状态，在保障论文不再重复分发推送的同时，也保障 iSwitch 系统分发推送的审计功能和异常处理能力，由于不确定的原因而导致论文没有正常推送的时候，系统会自动更新论文状态，保障任务代理再次调度此篇论文，最终分发到目的作者机构或者资助机构。

3.4 分发推送任务异常处理模块

iSwitch 系统分发任务标识设置了 4 个状态：1（可分发）、2（分发完成）、3（分发部分失败）和 4（分发全部失败）。当一个任务执行过程，出现论文因异常情况无法推送成功时，此任务会被标识为 3 或者 4（任务结束进行统计），此时系统仍然会自动调度此任务 2 次，如果仍然失败，则此任务不会被代理再次调度，需要人工审计，解决相应的问题之后，手动触发“分发推送任务异常处理模块”，完成未分发的论文的推送，如上图 2 所示。

4 iSwitch 论文分发推送审计模块设计

iSwitch 分发审计功能是基于分发任务的分发推送结果，审核与统计已经分发推送的论文概况，保障接收论文可靠分发，提升 iSwitch 平台的服务质量与效果，也是对出版商所提供数据的进一步检验。审计模块的设计从出版商、作者机构、资助基金以及分发任务 4 个维度，基于时间窗进行即时审计，审计各个维度需要分发推送的论文总数（根据 iSwitch 系统接收与解析论文的结果进行审计）、成功分发论文总数（分发状态为已分发）与失败分发论文总数（分发状态为失败）。

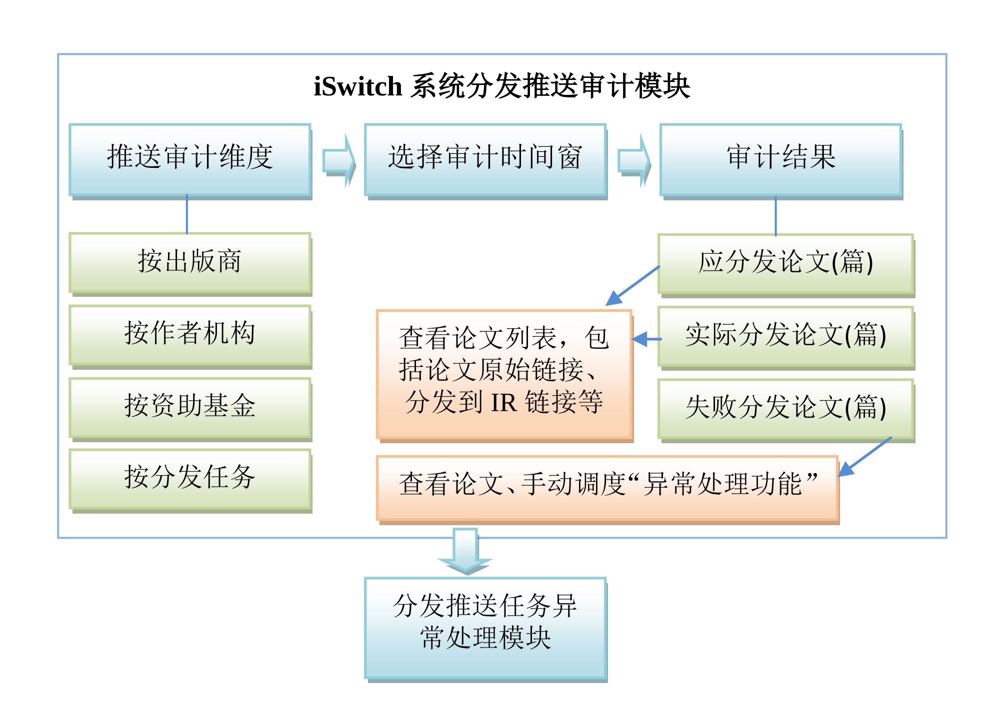


图 3　iSwitch 系统分发推送审计模块结构图

5 iSwitch 论文分发推送实验

为了验证 iSwitch 服务系统的论文推送转发服务的有效性，保障 iSwitch 系统在后期提供优质服务，笔者开展了相关实验，下面从实验准备条件和实现结果分析两个方面对分发服务机制进行介绍。

5.1 实验准备条件

开放获取论文推送者：汤森路透；数据来源：Web of Science；数据体量：34 332 条；覆盖中国科学院机构：99 个；覆盖资助基金：2 个；覆盖年限：2013 年；开放获取论文转发者：iSwitch；开放获取论文接受者：IR；转发通信协议：FTP。

5.2 实验结果与分析

（1）总体实验结果。通过定时调度 iSwitch 分发服务中的分发器，对 34 332 篇开放获取论文实验数据进行分发推送，详细统计结果如表 1 所示，实验结果整体概况如图 4 所示。结果表明，实验论文数据全部分发成功，而且一篇论文属于多个机构的情况也得到正确的分发推送到相对应的作者机构或者资助基金中，所以 iSwitch 服务系统的分发推送功能在分发结果上满足项目建设目标。

表 1　开放获取论文推送转发实验结果统计表

接收论文篇数	转发篇数	转发篇次数	失败转发篇数
34 332	30 339	56 299	0



图 4　iSwitch 服务系统实验分发推送结果图

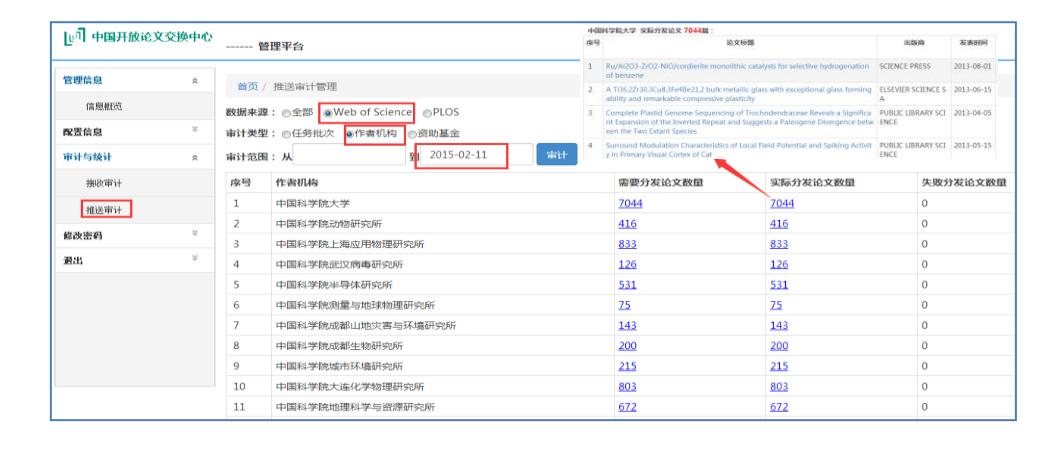
（2）**总体实验效率**。iSwitch 分发推送服务系统的分发通信协议支持 FTP 与 SWORD 两种，其中 FTP 在支持批量论文的转发方面效率更高，笔者在本次实验中即采用基于 FTP 通信协议，实验效率详细分析统计如表 2 所示，结果表明 iSwitch 推送转发服务系统基于 FTP 协议的转发服务效率很高，实现批量数据的及时转发到作者机构与资助基金的 IR 中，而作者机构可以快速收集本机构科研人员的科研成果。

表 2　开放获取论文推送转发实验分析统计表

分发论文总消耗时间（分）	平均转发每篇次消耗时间(秒/篇次)
12.5	0.013

（3）**实验审计结果**。从数据来源、审计类型作者机构三个维度对本次推送转发结果进行审计，结果如下图 5 所示，共有 99 个机构接收到推送转发的开放论文，审计结果显示，实验数据中涉及到的 99 个机构都正确接收到 iSwitch 转发推送的论文，进一步验证整个转发推送的工作流程是正确的。

图 5　iSwitch 服务系统实验分发推送审计图



6 结论

笔者按照 iSwitch 系统整体设计框架、技术规范以及分发推送流程实现了 iSwitch 系统分发推送功能，利用 FTP 通信协议方式，安全、高效地将分发论文数据分发推送到所对应机构的 FTP 目录下。iSwitch 系统分发推送功能所实现的“自动创建分发任务、自动执行分发任务与异常分发处理”机制保障了各个研究机构及时有效地收集该机构的论文资产，促进了机构知识库的建设和开放科技论文的传播与利用。后续研究工作将从以下两个方面重点展开：基于当前技术实现平台，开展基于多个数据来源、更大体量的开放科研论文的分发推送工作，根据实际分发推送效果对相关技术环节进一步完善；基于 iSwitch 分发推送服务，构建校验与完善机构知识库中论文信息的技术服务机制。

参考文献

[1] 张冬荣, 祝忠明, 李麟, 等. 中国科学院机构知识库建设推广与服务[J].图书情报工作, 2013, 57(1):20-25.（Zhang Dongrong, Zhu Zhongming, Li Lin, et al. Construction, Promotion and Service of CAS IRs[J]. Library and Information Service, 2013,57 (1): 20-25.）

[2] 张晓林, 刘细文, 李麟, 等. 研究图书馆推进开放获取的战略与实践——以国家科学图书馆为例[J]. 图书情报工作,2013, 57 (1): 15-19, 48. (Zhang Xiaolin, Liu Xiwen, Li Lin,et al. The Strategies and Practices of Research Library to Support Open Access——Taking National Science Library as an Example [J]. Library and Information Service, 2013,57 (1): 15-19, 48.)

[3]张晓林, 梁娜, 钱力, 等. 开放获取论文推送转发服务系统 iSwitch：概念、功能与基本框架[J]. 现代图书情报技术, 2014(10): 4-8. (Zhang Xiaolin, Liang Na, Qian Li, et al. Router Service Engine iSwitch for Open Access Articles: The Concept, Strategy, and Framework[J]. New Technology of Library and Information Service, 2014(10): 4-8.)

[4]梁娜, 张晓林, 钱力, 等. 开放获取论文推送转发服务系统 iSwitch：技术流程与标准[J]. 现代图书情报技术, 2014(10): 9-13. (Liang Na, Zhang Xiaolin, Qian Li, et al. Router Service Engine iSwitch for Open Access Articles: Technical Workflows and Standards[J]. New Technology of Library and Information Service, 2014 (10): 9-13.)

[5] Serv-U File Server Administrator Guide [EB/OL]. [2014-12-25]. http://www.serv-u.com/Serv-U-Administrator-Guide.pdf.

[6] Simple Web-service Offering Repository Deposit [EB/OL]. [2014-12-25]. http://swordapp.org/about/.

[7] Journal Article Versions (JAV): Recommendations of the NISO/ALPSP JAV [EB/OL]. [2014-12-25]. http://docs.rioxx.net/v2-0-beta-1/.

作者贡献说明

钱力：iSwitch 试验系统需求调研及论文分发部分的开发，撰写论文；

师洪波：iSwitch 试验系统需求调研及论文接收及解析部分的开发；

张晓林：提出和梳理技术流程及技术要求，审核论文；

梁娜：梳理提出技术流程和技术要求，提出参考技术标准。

(通讯作者：钱力，ORCID：0000-0002-0931-2882，E-mail:qianl@mail.las.ac.cn。)